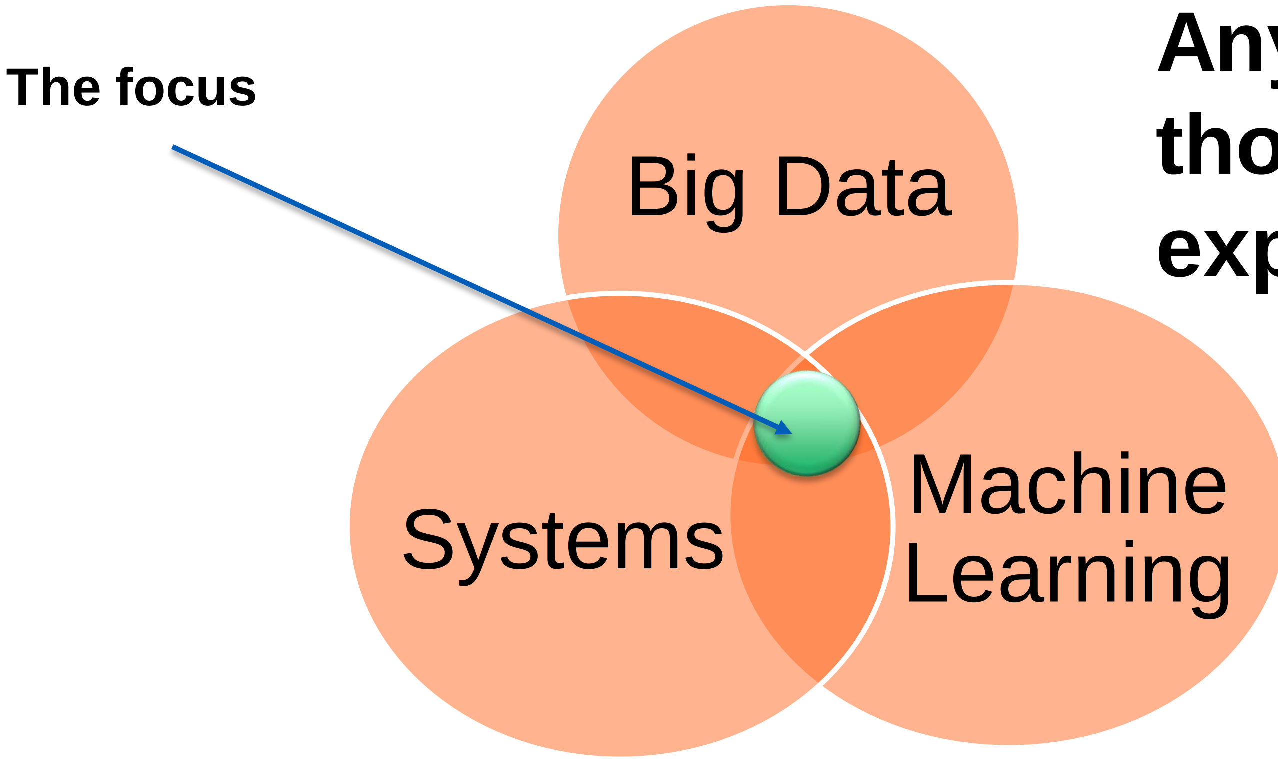
Aalto University
School of Science

Benchmarking, Monitoring and Validation for R3E

Hong-Linh Truong
Department of Computer Science
linh.truong@aalto.fi, *https://rdsea.github.io*

Our focus in this course



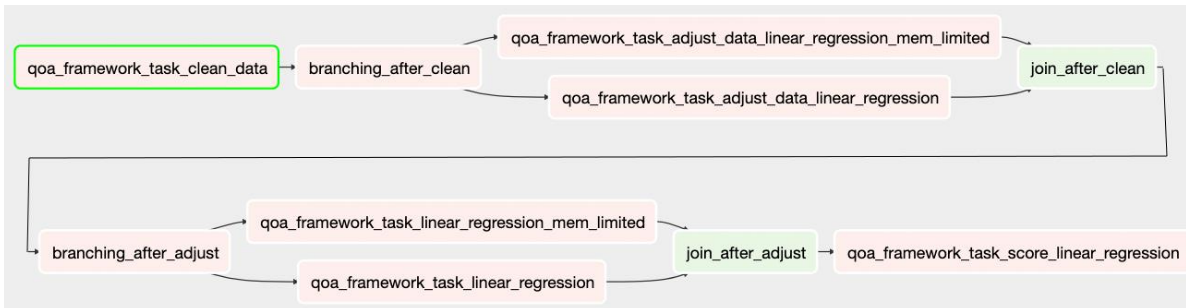
Any idea, thought, expectation?

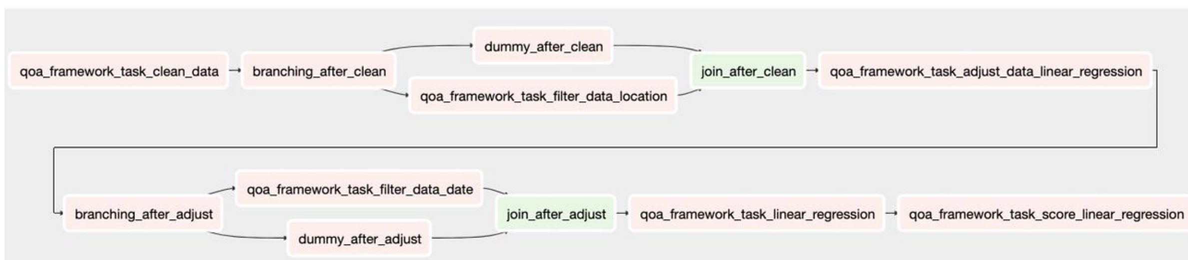
Content

- **Cases for brainstorming**
- **Metrics, benchmarking, and monitoring**
- **An example of the approach**
  - incident monitoring for big data

Use cases

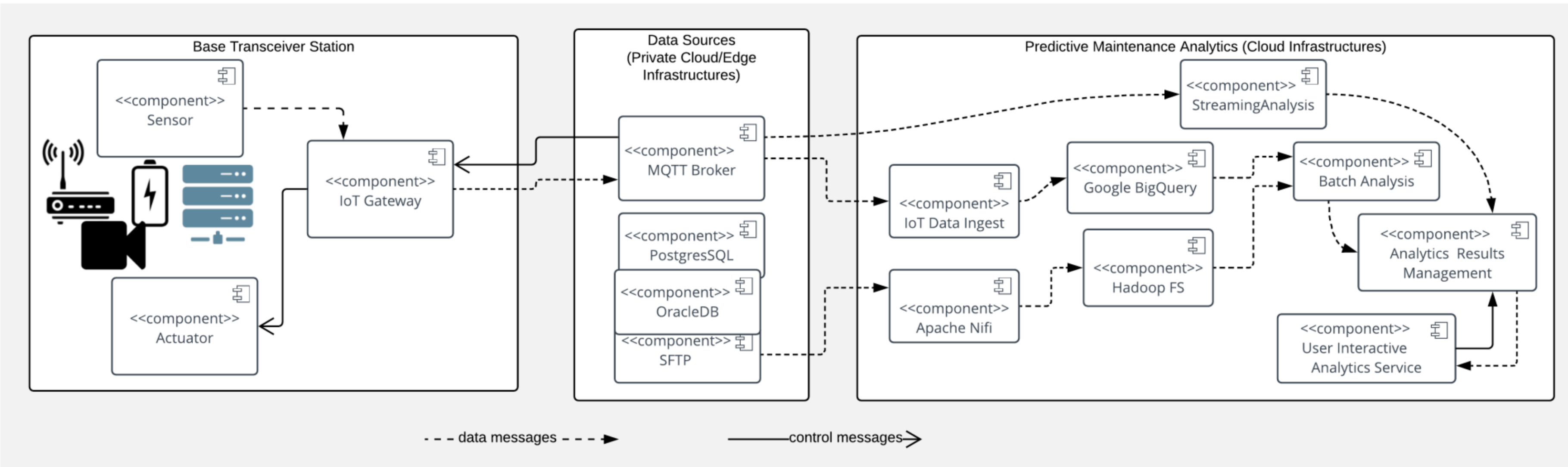
Discussion: what and how to monitor this retailing forecast pipeline





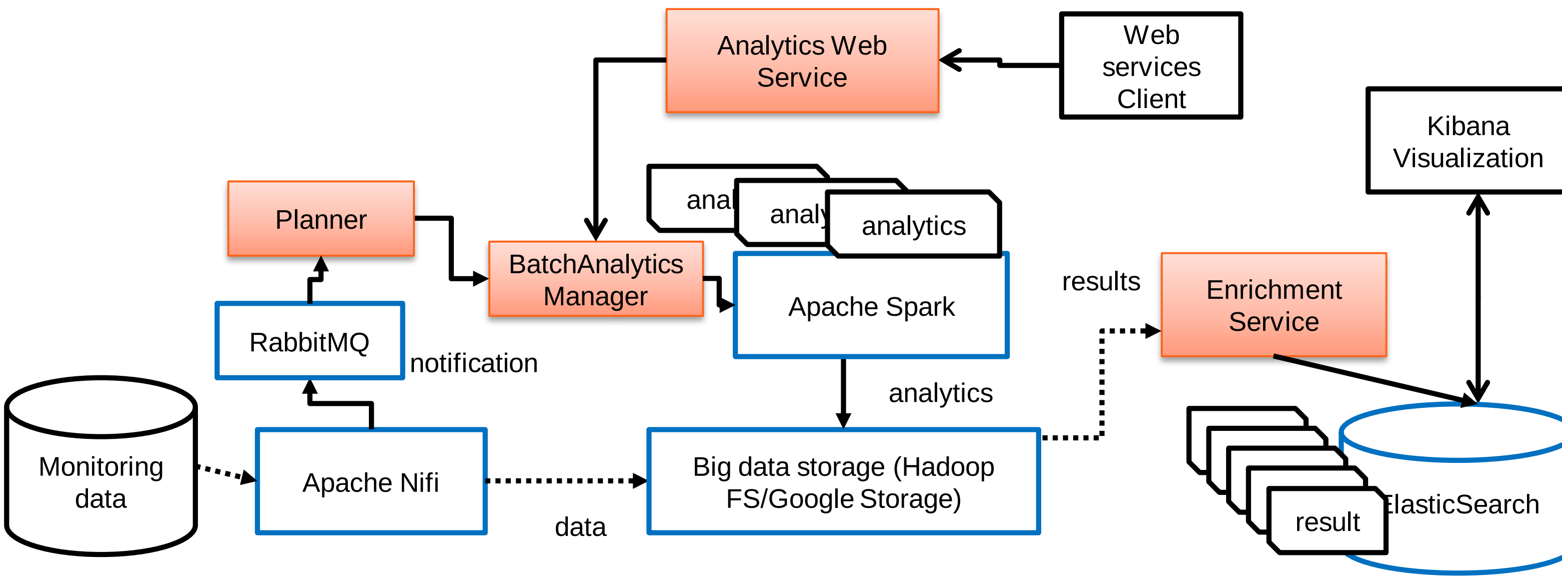
Source: Kreics Krists, „Quality of analytics management of data pipelines for retail forecasting,“, Aalto CS Master thesis, 2019

Discussion: what and how to monitor this big data platform



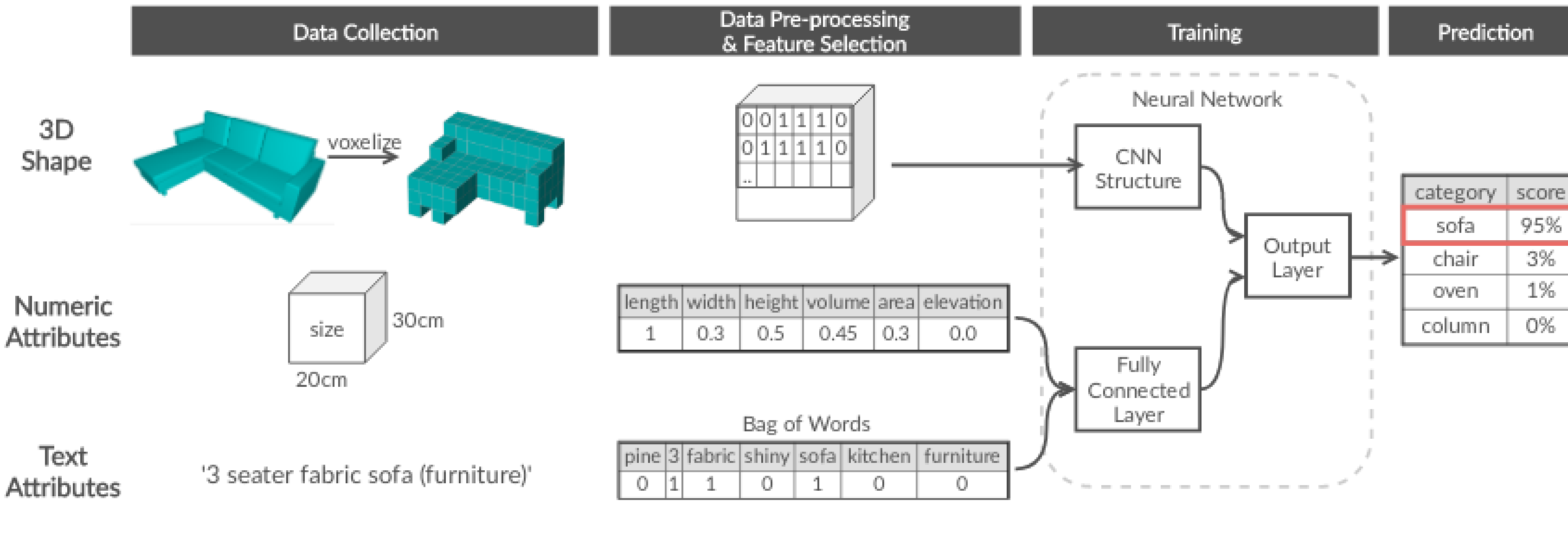
Source: Hong-Linh Truong, Integrated Analytics for IIoT Predictive Maintenance using IoT Big Data Cloud Systems, The IEEE International Conference on Industrial Internet (ICII 2018)

Discussion: what and how to monitor this analytics platform



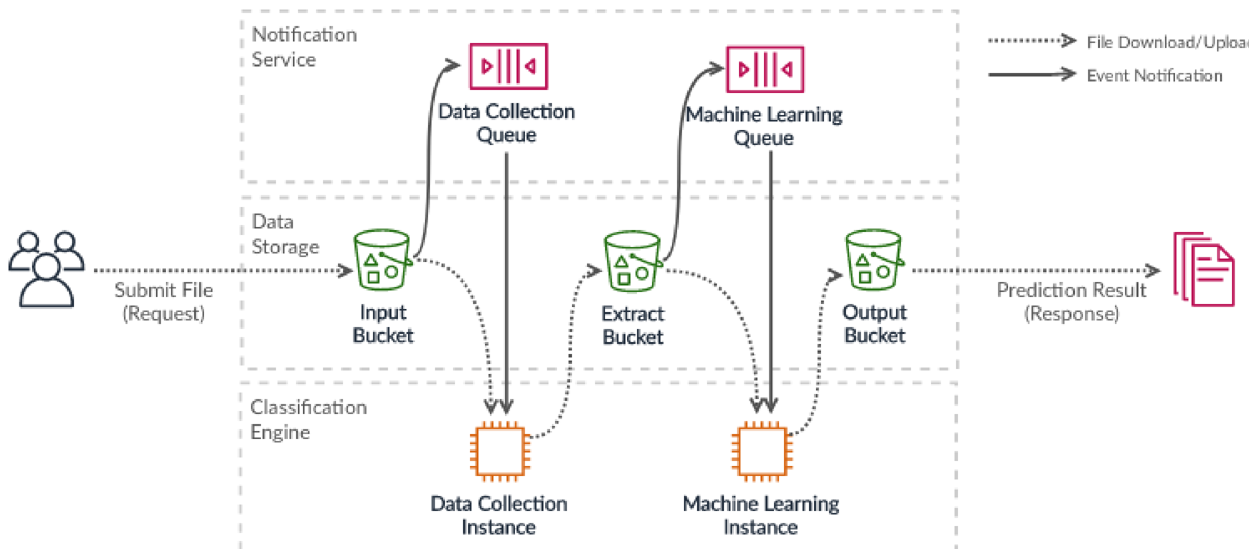
Source: Linh Truong, „*I & A Big Data Platform*“, Not published, 2018

Discussion: what and how to monitor this ML classification



Source: Minjung Ryu, „*Machine Learning-based Classification System for Building Information Models* “, Aalto CS Master thesis, 2020

Discussion: what and how to monitor this ML classification



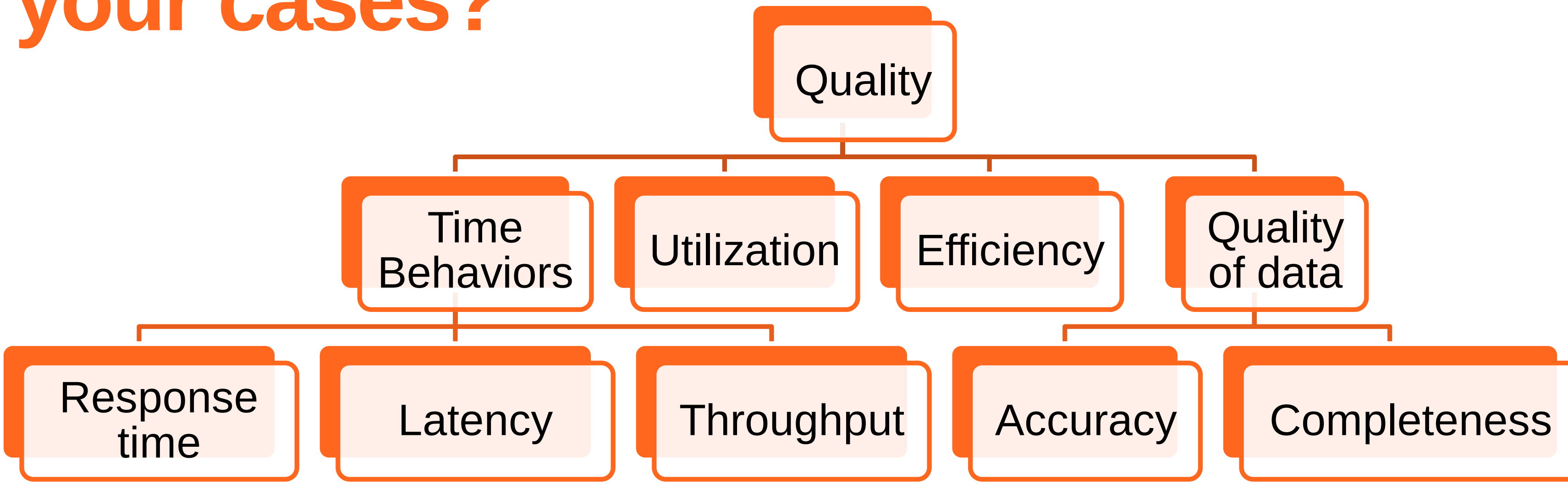
Source: Minjung Ryu, „*Machine Learning-based Classification System for Building Information Models* ", Aalto CS Master thesis, 2020

Methods

Key steps

- **Understand metrics characterizing big data/ML systems**
  - Common metrics but you might have some specific ones or have different relevance for your metrics
- **Monitoring, measurement and interpretation**
  - Understand dependencies among components
  - Understand tools for capturing metrics
  - Understand what kind of changes/designs we must do
  - Do monitor and analysis
  - Integrate many types of data for analytics

What are the most critical metrics for your cases?

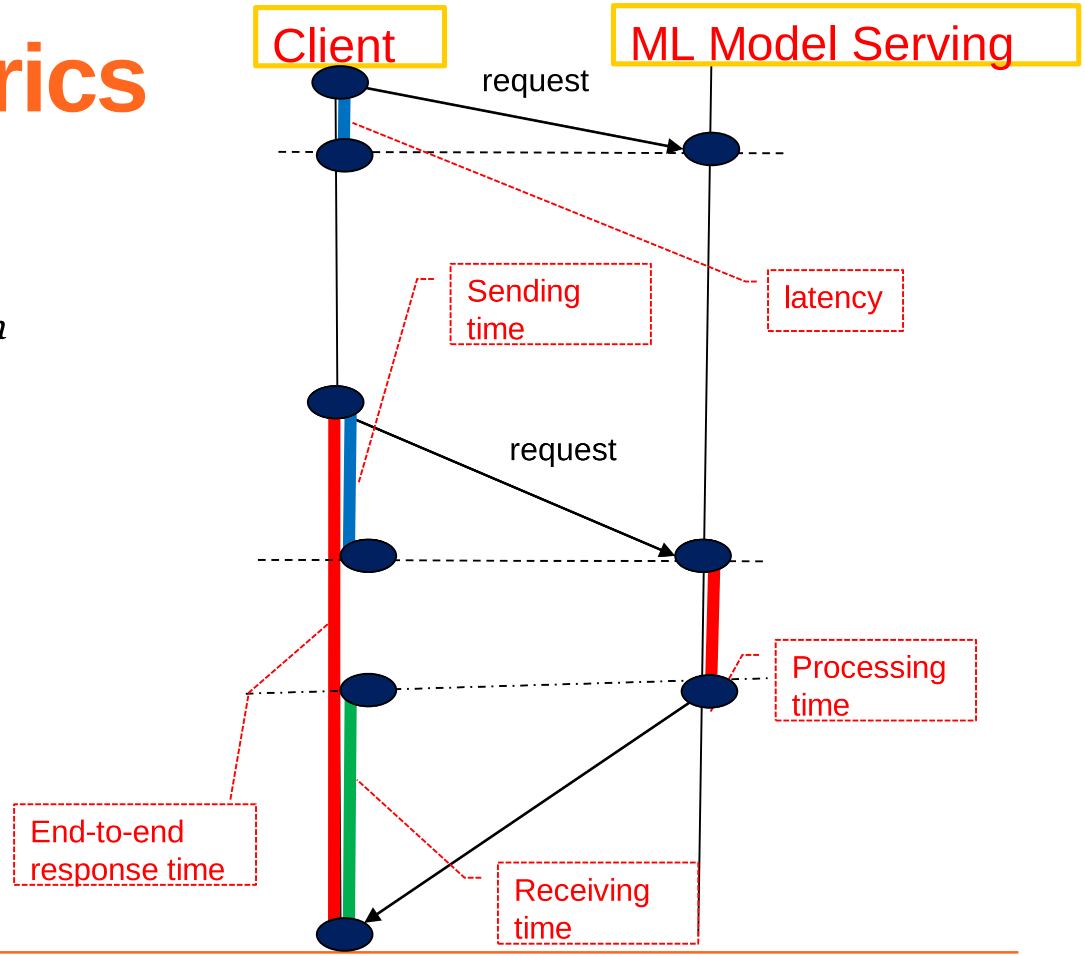


Industry view: https://guidingmetrics.com/content/cloud-services-industrys-10-most-critical-metrics/
NIST: https://www.nist.gov/sites/default/files/documents/itl/cloud/RATAX-CloudServiceMetricsDescription-DRAFT-20141111.pdf

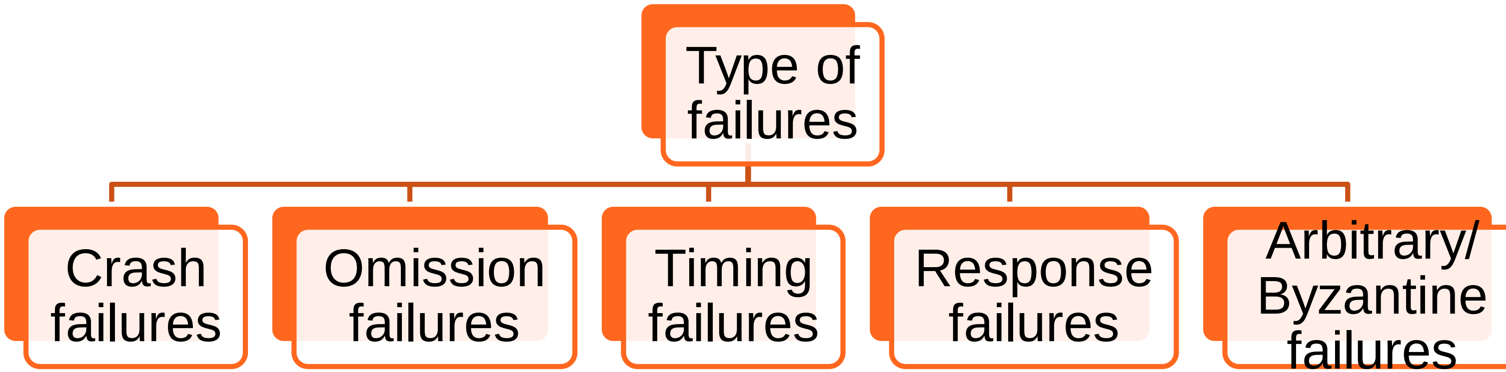
Common performance metrics

- **Timing behaviors**
  - Communication
    - *Latency/Transfer time*
    - *Data transfer rate, bandwidth*
  - Processing
    - *Response time*
    - *Throughput*
- **Utilization**
  - Network utilization
  - CPU utilization
  - Service utilization
- **Efficiency/Scalability**
  - Concurrent Executions

Examples



Failure


Type of failures
Crash failures
Omission failures
Timing failures
Response failures
Arbitrary/ Byzantine failures

Data Quality

- **Completeness**
- **Timeliness**
- **Currency**
- **Validity**
- **Format**
- **Accuracy**
- **Data Drift**

Metrics for ML models

- **Concept drift**
  - (https://en.wikipedia.org/wiki/Concept_drift)
- **Confusion matrix**
- **Accuracy**
- **Loss**
- **True positive rate**
- **False positive rate**
- **F1 Score/F-measure**
- **etc**

(see https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234)

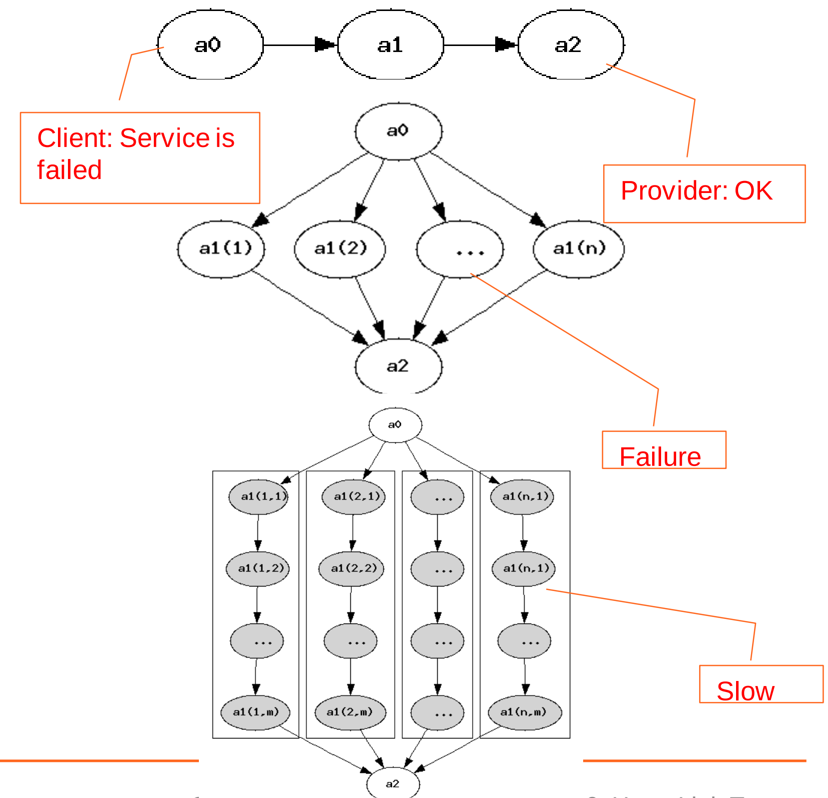
Dealing with benchmarking, monitoring, validation

- **Determines clearly system boundaries**
  - The system under study, the system used to judge, and the environment
- **Understands dependencies**
  - Among components in distributed big data/ML systems in distributed computing platforms
  - Single layer as well as cross-layered dependencies
- **Determines types of metrics and failures and break down problems along the dependency path**

Do we understand the structure of big data/ML application

- **Composable method**
  - Divide a complex structure into basic common structures
  - Each basic structure has different ways to analyze specific failures/metrics
- **Interpretation based on context/view**
  - Client view or service provider view?
  - Conformity versus specific requirement assessment

Dependency Structure



Support an end-to-end view or not?

- **What does it mean end-to-end? Examples?**
  - Reflect the entire system
  - E.g., data reliability: from sensors to the final analytics results
- **The user expects end-to-end R3E**
  - E.g., specified in the expected accuracy
- **Providers/operators want to guarantee end-to-end quality**
  - *Need to monitor different parts, each has subsystems/components*
  - *Coordination-aware assurance*
    - Elasticity principles?

Measurement, Monitoring and Interpretation

- **Instrumentation and Sampling**
  - Instrumentation: insert probes into systems so that you can measure system behaviors directly
  - Sampling: use components to take samples of system behaviors
- **Monitoring**
  - Probes or components perform sampling or measurements, storing and sharing measurements
- **Interpretation**
  - Evaluate and interpret measurements for specific contexts
  - Can be subjective!

Benchmarking

- **Big data pipelines**
  - Benchmark individual subsystems: brokers, databases, processing
  - Data ingestion throughput, processing throughput and time, component CPU and memory
- **ML pipelines**

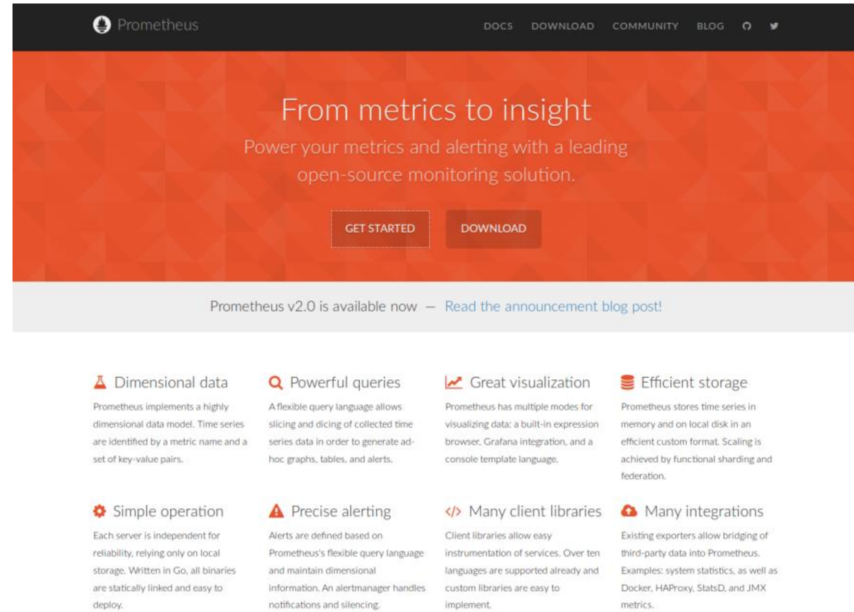
Benchmark	Dataset	Quality Target	Reference Implementation Model
Image classification	ImageNet (224x224)	75.9% Top-1 Accuracy	Resnet-50 v1.5
Object detection (light weight)	COCO 2017	23% mAP	SSD-ResNet34
Object detection (heavy weight)	COCO 2017	0.377 Box min AP, 0.339 Mask min AP	Mask R-CNN
Translation (recurrent)	WMT English-German	24.0 BLEU	GMNT
Translation (non-recurrent)	WMT English-German	25.0 BLEU	Transformer
Recommendation	Undergoing modification		
Reinforcement learning	N/A	Pre-trained checkpoint	Mini Go

Source: https://mlperf.org/training-overview

System Monitoring Tools

There are many powerful tools!

But only system information (infrastructures)



From: https://prometheus.io/

Instrumentation

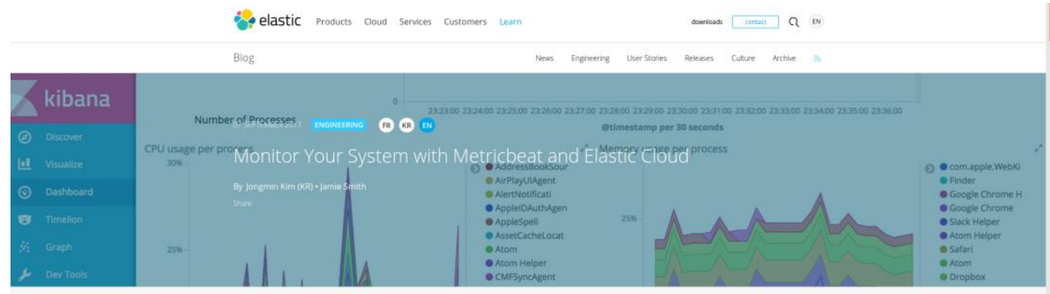
Code instrumentation and logs

From: https://www.fluentd.org/

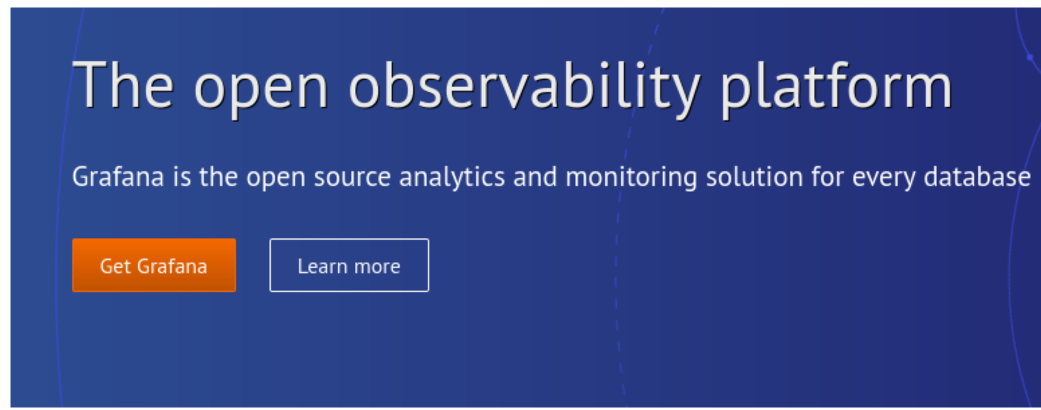
Visualization

Metrics and Visualization

- Easy to visualize many types of metrics
- But only you can specify, define and map to your applications



https://www.elastic.co/products/kibana



https://grafana.com/

Data & Model Validation/Analysis

- **By humans or by software?**
- **Data validation tools are very diverse, depending on the frameworks and data**
  - E.g., Tensors Flows: https://www.tensorflow.org/tfx/guide/tfdv
- **Model Analysis:**
  - E.g., https://www.tensorflow.org/tfx/model_analysis/

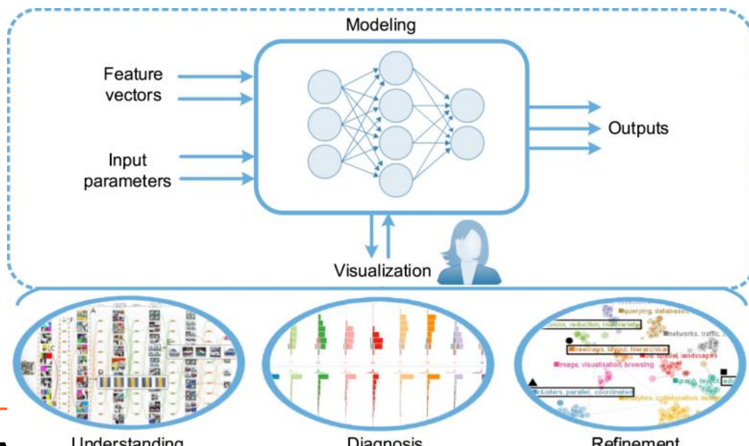


Figure source: Shixia Liu, Xiting Wang, Mengchen Liu, Jun Zhu,
Towards better analysis of machine learning models: A visual analytics perspective,
Visual Informatics, Volume 1, Issue 1, 2017,
https://doi.org/10.1016/j.visinf.2017.01.006.

Examples

Incidents in cloud-based big data

If you monitor alarms in a station and see this

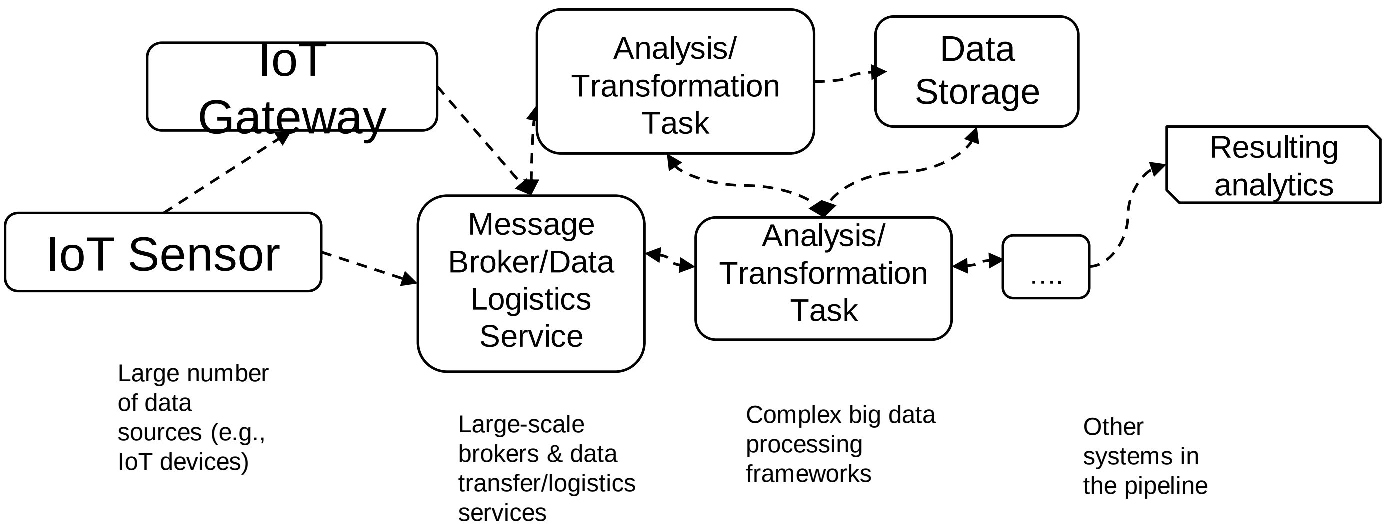
What could be happened?

Steps: Incident monitoring and analytics

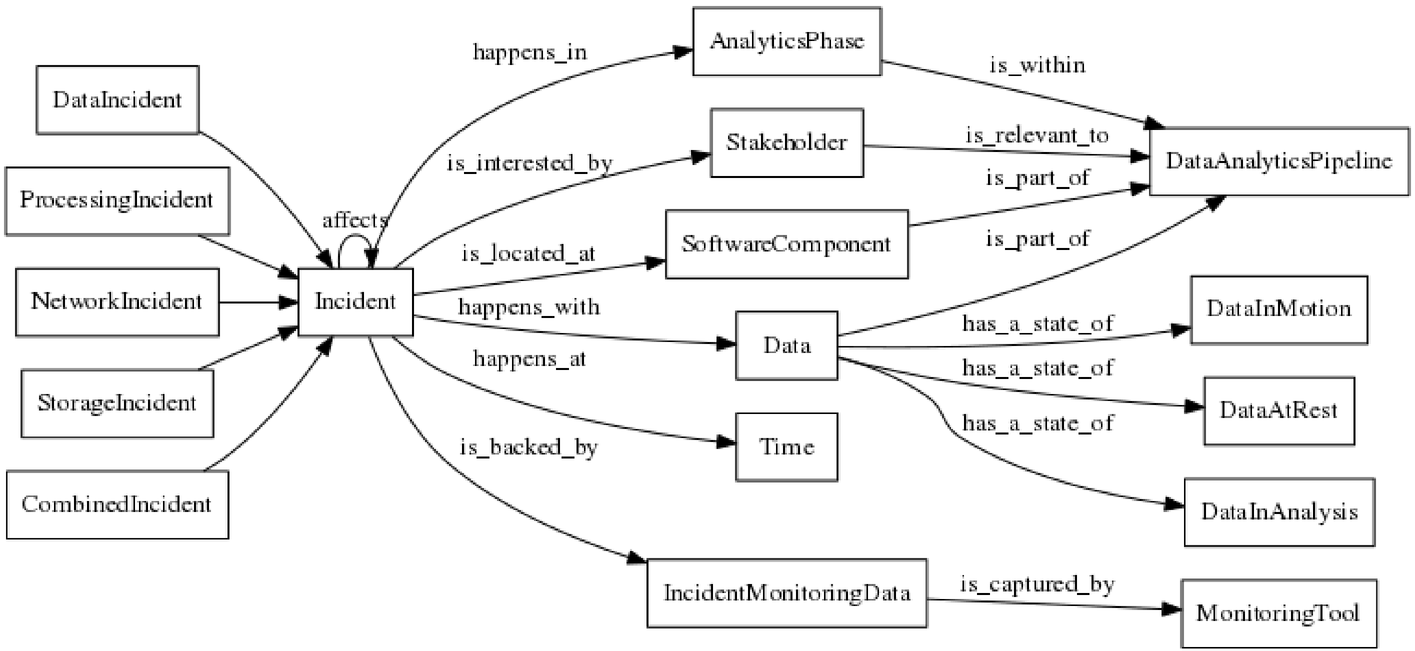
- **Classification of incidents:**
  - to quantify incidents and identify possible data sources, monitoring techniques and analytics.
- **Measurement/Instrumentation:**
  - to provide mechanisms for measurement and data collection for incidents.
- **Incident analytics:**
  - to find out the root cause and dependencies of incidents.

W3H: what, when, where and how for incidents

Too complex with many types of software. Can we have a simplified taxonomy for mapping incidents?



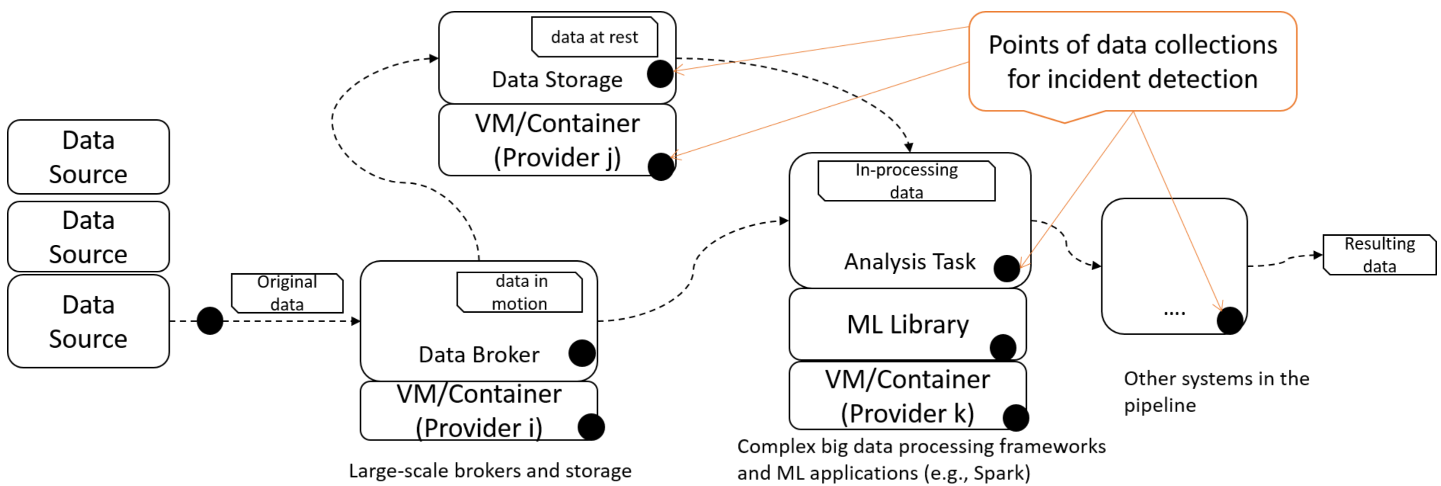
Examples: classification of incidents



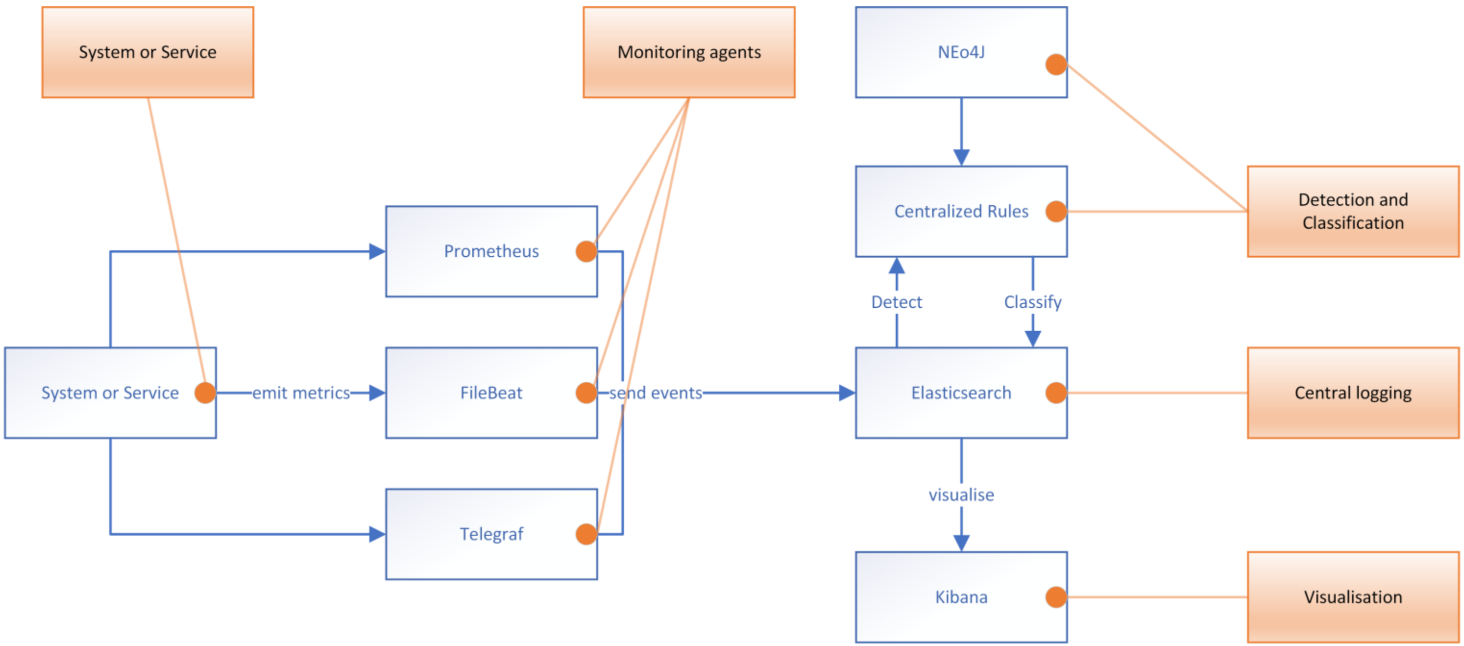
Hong-Linh Truong, Manfred Halper, **Characterizing Incidents in Cloud-based IoT Data Analytics**,, The 42nd IEEE International Conference on Computers, Software & Applications Tokyo, Japan, July 23-27, 2018.

Points of instrumentation for gathering data for incident analytics

Capture monitoring data to analyze and solve incidents, especially incidents related to data quality, across subsystems in ensembles to achieve quality of results



Integration monitoring and instrumentation



But who are going to define metrics and metric analysis?

Study log 2

Describe one big data/ML pipeline that you are familiar with and explain your thoughts on how would you support "benchmarking", "monitoring" or "validation" for testing/implementing R3E aspects

- Is enough to focus on 1 pipeline and 1 aspect
- Be concrete, e.g., with metrics and possible tools
- Analyze if things can be done easily or where are the challenges that might be interesting for further investigation

Thanks!

Hong-Linh Truong
Department of Computer Science

rdsea.github.io